BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software
Volume 8(25)

ExonVisualiser – application for visualization exon units in 2D and 3D protein structures

Monika Piwowar[1]*, Porembski Krzysztof[1] & Piwowar Piotr[2]

[1]Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Lazarza 16, 31-530 Krakow, Poland; [2]Department of Measurement and Electronics, AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland; Monika Piwowar - Email: mpiwowar@cm-uj.krakow.pl; *Corresponding author

Received November 12, 2012; Accepted November 14, 2012; Published December 19, 2012

Abstract:
The web application oriented on identification and visualization of protein regions encoded by exons is presented. The Exon Visualiser can be used for visualisation on different levels of protein structure: at the primary (sequence) level and secondary structures level, as well as at the level of tertiary protein structure. The programme is suitable for processing data for all genes which have protein expressions deposited in the PDB database. The procedure steps implemented in the application: I) loading exons sequences and theirs coordinates from GenBank file as well as protein sequences: CDS from GenBank and aminoacid sequence from PDB II) consensus sequence creation (comparing amino acid sequences form PDB file with the CDS sequence from GenBank file) III) matching exon coordinates IV) visualisation in 2D and 3D protein structures. Presented web-tool among others provides the color-coded graphical display of protein sequences and chains in three dimensional protein structures which are correlated with the corresponding exons.

Availability: http://149.156.12.53/ExonVisualiser/

Keywords: Exon visualisation, Exon unit identification in proteins

Background:

The presented aplication has been developed due to the lack of efficiently working tools for identification and visualisation of exons in the structures of proteins, as well as due to lack of information about exons in protein databases **[1-7].** Filling up the existing gap seemed recommended so as to enable tracking of the process of expression of the basic genetic information in the form of gene splitting into exons unit in proteins. One of many commonly available web browsers is necessary for using the application. The tests have been conducted for the browsers in the following versions: Firefox 4, Opera 11, Internet Explorer 8, Chrome 11, and Safari 5. In case of older versions of browsers, minor visual differences may occur caused with differences in the method of interpretation of GUI application styles, yet these differences should not affect its functionality.

Methodology:

The process leading to visualisation of exons in protein structure, implemented in the ExonVisualiser application, runs in two stages. The first part (identification of exons in the amino acid sequence of protein) is much more critical than the second part (visualisation of the exons found in the 2D and 3D structure of protein).

Identification of exons in protein structures:

Loading input data

The first step of the identification algorithm is loading of the data provided by the user which are related to both the nucleotide sequence (the GenBank file) and protein structure (the PDB file); GenBank file parts crucial from the point of view of tracing exons: The nucleotide sequence; CDS covering only the coding region; The splitting into exons; (In case of the ExonVisualiser application, it has been assumed that lack of information on exons means that one exon creates the entirety of the sequence). Basic header information and Accession Number; The data loaded from the PDB file include, in the first order: The protein amino acid sequence readied on the basis of the ATOM section; If more than one model is found in the PDB file, only the first model is extracted in the algorithm used in the created application is considered to be the default one. The HELIX and SHEET section which stores the data's on the secondary structures present in the protein, the helixes and beta structures, respectively. Apart from the structure type, the range of indexes is loaded which define the amino acids included in the given particle of protein. The PDB file header with the identifier of the given structure.

Matching the sequence

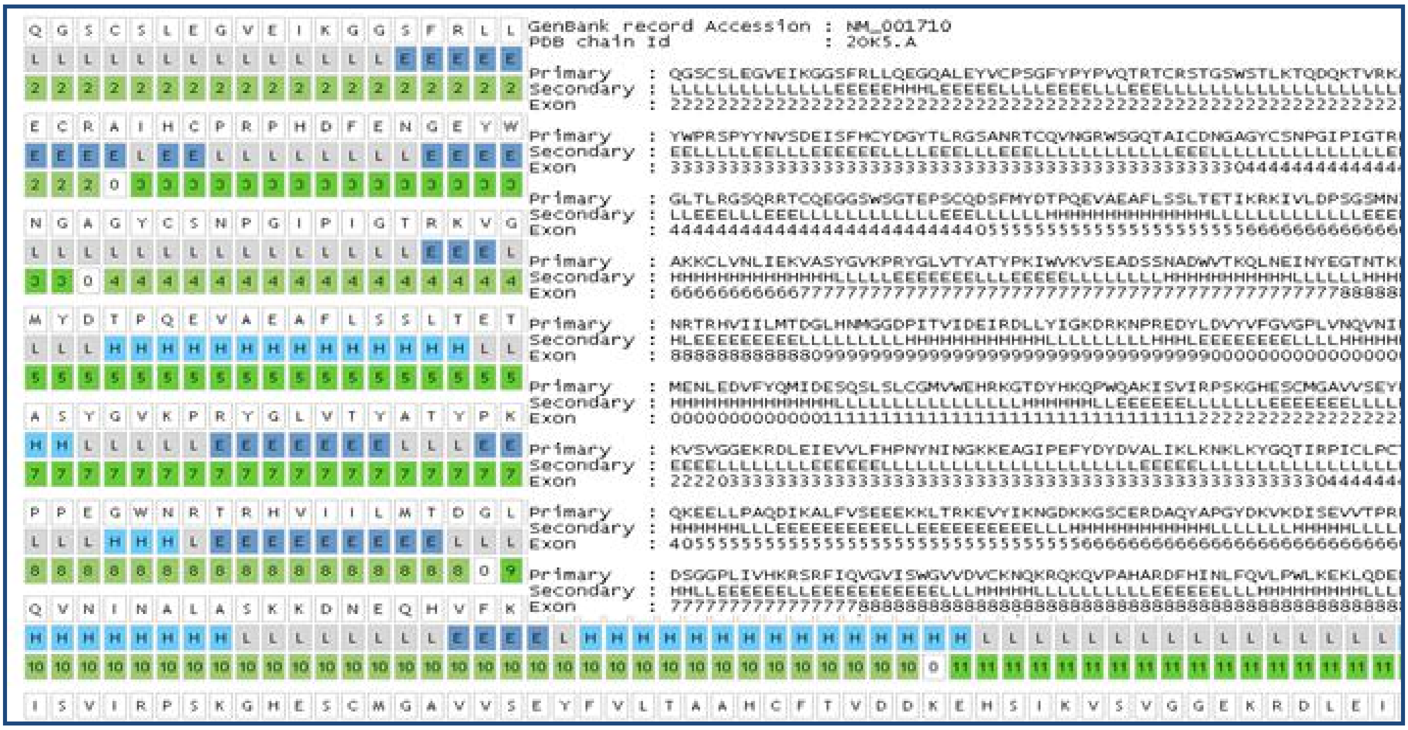
With comparable amino acid sequences (the CDS sequence which theoretically should develop on the basis of the given RNA, and the experimentally obtained sequence of protein from the PDB file), their comparison may be initiated. In the issue described, comparison of sequences pairs is employed. To be exact, it is crucial to find the consensus sequence, which is most probably the common unmodified part of the gene sequence translated into the amino acid language with the sequence of functional protein resulting from it. Creating such a sequenceenables further analysis for identification of exons in protein structures. Thus, matching the sequence is a key step from the point of view of the problem at hand. It is interesting to note the assumption in the issue here that the sequence of the analysed protein is coming from the analysed gene. In the ExonVisualiser application, matching sequence is done with the Blast programme **[8].** In the developed application, in which finding exons in the protein is the point, the amino acid sequence of protein from PDB is the source sequence (*subjects*)in the Blast programme. The transcript sequence given as *query* is matched with it.

Matching exons

In the next step, finding the consensus sequence in the source sequence resulting from the comparison is necessary (the sequence from the PDB file).With a similar process run for the matched sequence (the data from the GenBank file), identity of the individual amino acids from the consensus sequence to exons may be determined. Combining the two above sets of data, the information is obtained about the identity of the individual amino acids in the protein to exons. Thus, the first objective of the algorithm is attained: identification of exons in the protein sequence. In the developed application, in which finding exons in the protein is the point, the amino acid sequence of protein from PDB is the source sequence (*subjects).*The transcript sequence given as *query* is matched with it.

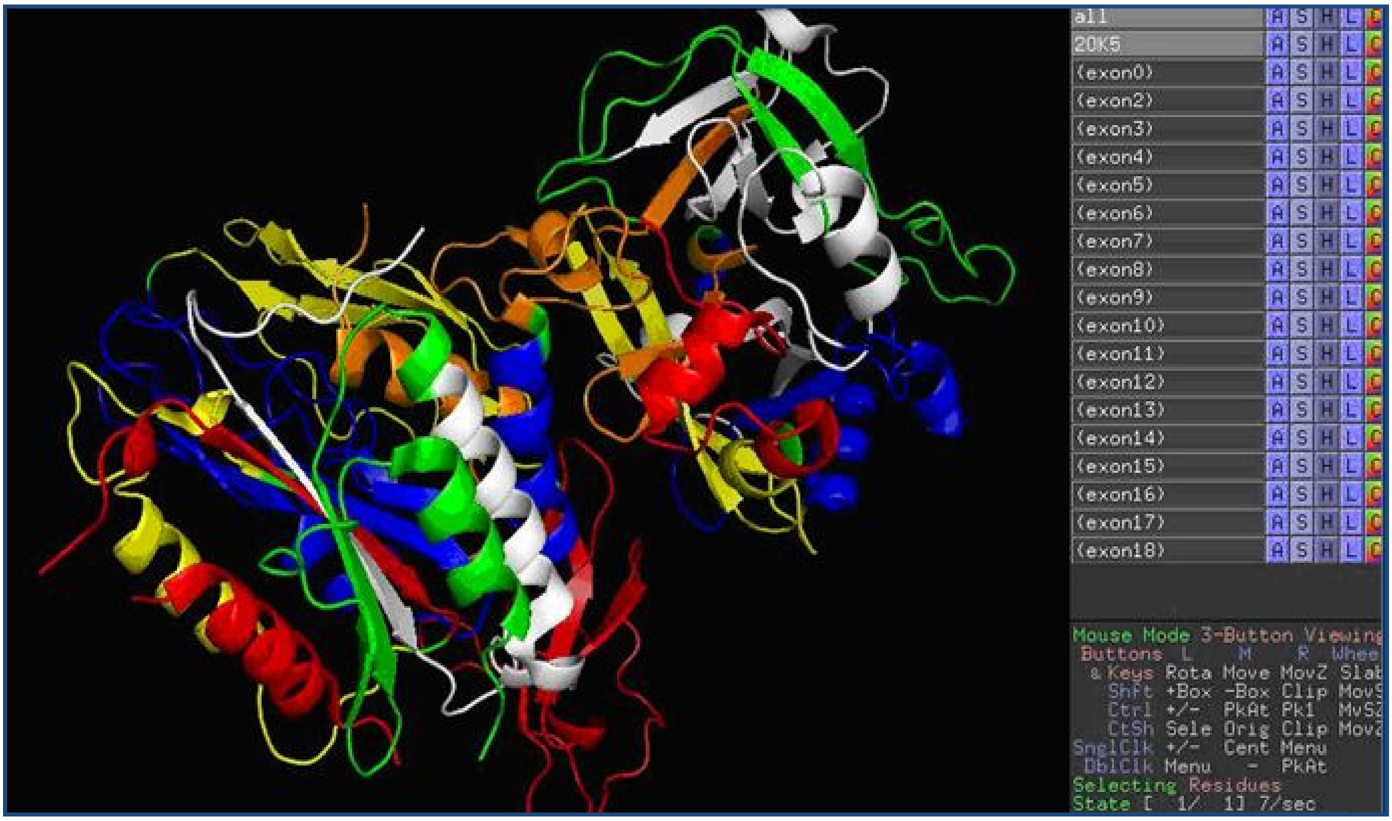
Identification of exons in protein structures

The data on protein secondary structures are loaded along with the information about its amino acid sequences from the PDB file. Each structure has closely defined limits in this sequence. After identification of exons in the primary structure, transferring the information about them on the secondary structure is a minor thing. Transferring the results of exon identification in the protein sequence on the level of the tertiary structure is similar. This operation is not a problem, as the protein sequence is determined on the basis of the ATOM section from the PDB file and there is full analogy between it and the tertiary structure.



```
GenBank record Accession : NM_001710
PDB chain Id             : 2OK5.A

Primary   : QGSCSLEGVEIKGGSFRLLQEGQALEYVCPSGFYPYPVQTRTCRSTGSWSTLKTQDQKTVRK
Secondary : LLLLLLLLLLLLLLLEEEEEHHHLEEEEELLLLEEEEELLLEEELLLLLLLLLLLLLLLLLLL
Exon      : 2222222222222222222222222222222222222222222222222222222222222222

Primary   : YWPRSPYYNVSDEISFHCYDGYTLRGSANRTCQVNGRWSGQTAICDNGAGYCSNPGIPIGTR
Secondary : EELLLLLLEELLLEEEEEEELLLLEEELLLEEELLLLLLLLLLLLEEELLLLLLLLLLLLLLLE
Exon      : 3333333333333333333333333333333333333333333333333330444444444444

Primary   : GLTLRGSQRRTCQEGGSWSGTEPSCQDSFMYDTPQEVAEAFLSSLTETIKRKIVLDPSGSMN
Secondary : LLEEELLLEEELLLLLLLLLLLLLEEELLLLLLLHHHHHHHHHHHHHHLLLLLLLLLLLLLLEEE
Exon      : 4444444444444444444444444405555555555555555555555555666666666666

Primary   : AKKCLVNLIEKVASYGVKPRYGLVTYATYPKIWVKVSEADSSNADWVTKQLNEINYEGTNTK
Secondary : HHHHHHHHHHHHHHHLLLLLEEEEEEEELLLEEEEELLLLLLLLHHHHHHHHHHHLLLLLLHHH
Exon      : 6666666666677777777777777777777777777777777777777777777777788888

Primary   : NRTRHVIILMTDGLHNMGGDPITVIDEIRDLLYIGKDRKNPREDYLDVYVFGVGPLVNQVNI
Secondary : HLEEEEEEEEEELLLLLLLLLHHHHHHHHHHHHHLLLLLLLLLLHHHLEEEEEEEEELLLLHHHHH
Exon      : 8888888888880999999999999999999999999999999999999900000000000000

Primary   : MENLEDVFYQMIDESQSLSLCGMVWEHRKGTDYHKQPWQAKISVIRPSKGHESCMGAVVSEY
Secondary : HHHHHHHHHHHHHHHLLLLLLLLLLLLLLLLLHHHHHHLLEEEEEEELLLLLLLEEEEEEEEELLLL
Exon      : 0000000000000011111111111111111111111111111111111122222222222222

Primary   : KVSVGGEKRDLEIEVVLFHPNYNINGKKEAGIPEFYDYDVALIKLKNKLKYGQTIRPICLPC
Secondary : EEEELLLLLLLLLEEEEEEELLLLLLLLLLLLLLLLLLLLLLLLLLEEEEEELLLLLLLLLLLLLLLL
Exon      : 2222033333333333333333333333333333333333333333333333333304444444

Primary   : QKEELLPAQDIKALFVSEEEKKLTRKEVYIKNGDKKGSCERDAQYAPGYDKVKDISEVVTPR
Secondary : HHHHHHLLLEEEEEEEEEEELLEEEEEEEEEEELLLHHHHHHHHHHHHHLLLLLLLHHHHHLLLL
Exon      : 4055555555555555555555555555555555555556666666666666666666666666

Primary   : DSGGPLIVHKRSRFIQVGVISWGVVDVCKNQKRQKQVPAHARDFHINLFQVLPWLKEKLQDE
Secondary : HHLLEEEEEELLEEEEEEEEEEEELLLHHHHHLLLLLLLLLEEEEEELLLHHHHHHHHHLLL
Exon      : 7777777777777777888888888888888888888888888888888888888888888888
```

Figure 1: Identification of exons covering amino acid sequences (Html and text format) (Protein ID: 2K05; mRNA ID: NM_001710) Primary (first line) – amino acid sequence, Secondary (second line) – secondary structure identifier, Exon (third line) – exon identifier.



```
Mouse Mode 3-Button Viewing
 Buttons  L    M    R  Whee
  & Keys Rota Move MovZ Slab
    Shft +Box -Box Clip MovS
    Ctrl +/-  PkAt Pk1  MvSZ
    CtSh Sele Orig Clip MovZ
 SnglClk +/-  Cent Menu
  DblClk Menu  -   PkAt
Selecting Residues
State [  1/   1] 7/sec
```

Figure 2: Exons visualisation covering protein structure (structure ID: 2K05; mRNA ID: NM_001710) in PayMOL; exon0 – amino acid originates from two different exons; exson1is cut in the process of post-translational modifications thus does not appear in the functional protein; exon2 - exon18 – exons coded 2K05 protein

Visualisation of exons in protein structures

Visualisation is done in two ways: 1) the sequences and the secondary structures level and 2) the tertiary protein structure. The successive amino acids are presented linearly along with the information about the position in the sequence, the secondary structure to which they belong, the exon **(Figure 1).** In case of the tertiary structure, visualisation must be done with the application used for presentation of 3D structures and one of the formats supported by such a tool **(Figure 2).** Due to the origin of the protein structural data from the PDB file, the format is preferred for the input data. However, this does not enable appending of the result data of the developed application so that they could be unanimously read by different programmes used for particle visualisation. It is possible to report the results from the script written for specific software, which includes commands specific for it. The results may be presented in three ways: the default method, always developed, is preparing the view to be presented in an HTML page. Additionally, it is possible to have the results in the form of a text file and in the form of a PDB file. The results in these two formats may be downloaded with the appropriate option selected from the basic view menu (more information at "*Documentation*"http://149.156.12.53/ExonVisualiser/).

References:

[1] Shepelev V & Fedorov A, *Brief Bioinform*. 2006 **7:** 178 PMID: 16772261]

[2] Sakharkar M *et al. Nucleic Acids Res*. 2002 **30**: 191 [PMID: 11752290]

[3] Bhasi A *et al. Nucleic Acids Res*. 2009 **37**: D703 [PMID: 18984624]

[4] Gouzy J *et al. Comput Appl Biosci*. 1997 **13**: 601 [PMID: 9475988]

[5] Medvedeva I *et al. Nucleic Acids Res*. 2012 **40**: D278 [PMID: 22139920]

[6] Leslin CM *et al. Bioinformatics*. 2004 **20**: 1801 [PMID: 14988102]

[7] Stamm, S *et al. DNA and Cell Biol*. 2000 **19**: 739 [PMID: 11177572]

[8] Altschul SF *et al. J Mol Biol*. 1990 **215**: 403 [PMID: 2231712]

Edited by P Kangueane

Citation: Piwowar *et al.* Bioinformation 8(25): 1280-1282 (2012)